Exploratory Data Analysis of the Apple Quality Dataset

Author: Somraj Bharadwaj Cheppela
Student Id: 23032481
GitHub Repository: SomrajBharadwaj/Apple-Quality (github.com)

1. Introduction

In this report, I will perform an Explanatory Data Analysis (EDA) on Apple Quality sourced from Kaggle. This dataset contains information about attributes of a set of fruits providing their characteristics. The dataset includes details such as fruit ID, size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and quality.

2. Dataset Overview

Let's start by loading the dataset and exploring its structure. It consists of 4000 rows and 9 columns. This dataset consists of both numerical and categorical data. The numerical features include attributes such as fruit ID, size, weight, sweetness, crunchiness, juiciness, ripeness, and acidity while the categorical feature is the quality Depending on different attributes the quality of apples is classified as good or bad.

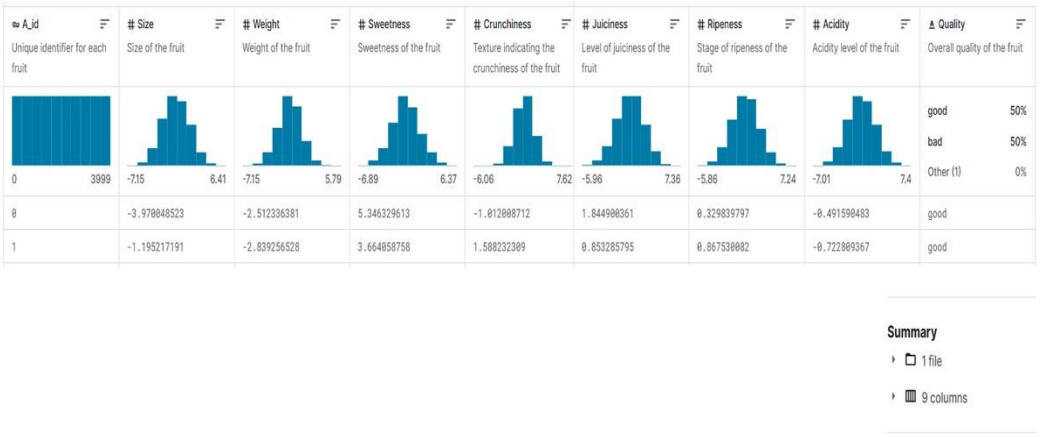

Fig 1: Overview of the Dataset

3. Exploratory Data Analysis:

Dataset Exploration:

I started the analysis by exploring the dataset. The dataset comprises eight numerical features and one categorical feature. The numerical features include attributes such as fruit ID, size, weight, sweetness, crunchiness, juiciness, ripeness, and acidity while the categorical feature is the quality. For getting a correlation among the attributes the

categorical feature of quality is replaced with good as ‘1’ and bad as ‘0’. Then we can have all attributes in numerical values.

4. Visualization of Graphs:

Histogram:

I visualized the distribution of Apple Quality using the histogram. The distribution is about the Sweetness of apples from the dataset. A histogram is used for visualizing the interpretation of numerical data. In this histogram, the x-axis displays the sweetness and the y-axis shows the count of given data. Depending on the values of the data, the histogram takes a different shape.

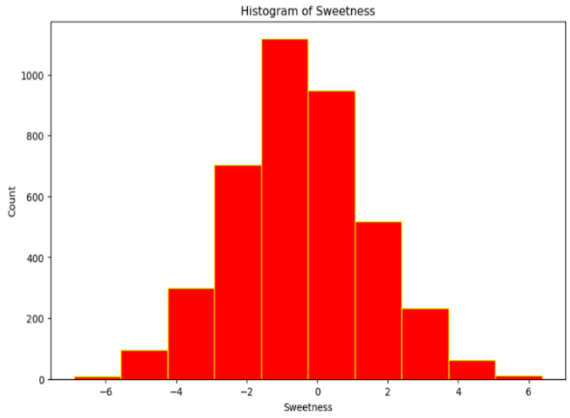

Fig 2: Histogram of Sweetness

Scatter Graph:

The goal of this scatter graph is to examine the relation between the size and weight of the apple. The variables “size” and “weight” are taken on the x-axis and the y-axis. Points on the graph showcase the relationship between the Size and Weight attribute of apple quality as shown.

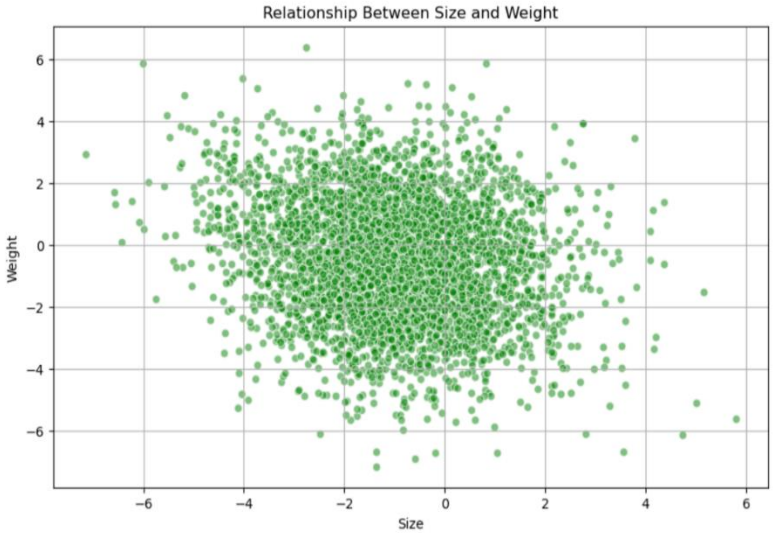

Fig 3: Scatter graph

Heatmap:

The objective of this report is to analyze the relation between various attributes of fruits, providing insights into their characteristics. By using the correlation between all the attributes of the fruits a correlation matrix is made. With the help of the correlation matrix heatmap is visualized. I aim to identify any patterns or discrepancies in the quality of apples using different attributes.

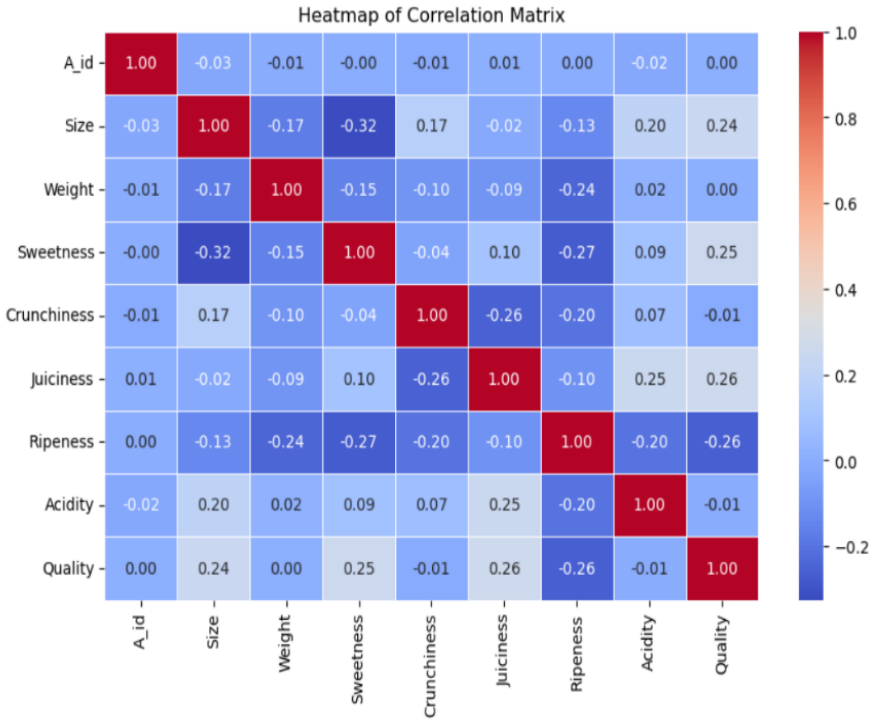

Fig 4: Heatmap

5. Conclusion

In this report, I loaded the dataset, exploded its structure, and computed summary statistics. Analyzing the dataset on Apple Quality provides valuable insights into the different relationships between attributes of Apple quality. Also, I visualized the sweetness of the fruit, compared the size and weight of the apple, and examined the correlation between the attributes of apple quality. The visualizations helped us understand patterns and distributions within the dataset, while descriptive statistics and correlation analysis quantified the relationships between numerical features. Further analysis could offer deeper insights into the dynamics of the quality of apples and their contributions to the health of an individual.